UDC 551.509.314:519.2

Conditional Probability for an Exact, Noncategorized Initial Condition

IRVING I. GRINGORTEN—*Air Force Cambridge Research Laboratories, Bedford, Mass.*

ABSTRACT—Previous models for estimating the conditional probability of an event have used, as the condition, an initial categorized event such as *no rain* or *overcast* at time zero. But initial conditions frequently are observed and known in greater detail, and these observed values can replace the categories in determining conditional probabilities. A model that has as its underlying assumption the "Ornstein-Uhlenbeck" process is applicable to this problem. It uses the antecedent quantitatively without loss of information and with surprising simplicity.

1. INTRODUCTION

In a recent report on modeling conditional probability, Gringorten (1971) obtained an estimate of the conditional probability of an event through use of the bivariate normal distribution, which is based on the unconditional climatic frequencies and persistence of the event. The assumption that the process is Markov is considered valid, and the estimates should be effective. However, usefulness of the method is curtailed by the fact that the initial condition is a category, such as an overcast sky, all temperatures less than 32°F, or 24-hr rainfall exceeding 0.1 in. At any initial time, however, an event such as the temperature is known specifically as, for example, 15°F. Its classification into a category of all instances of temperature equal to or less than 32°F results in loss of information. It is possible to use the specific information of initial state to yield a sharper conditional probability than that for the categorized initial state. Surprisingly, the solution proposed in this paper is simpler than the previous solution.

2. MODELING FOR A SPECIFIC INITIAL VALUE

The cumulative frequency distribution of a meteorological element, T, can be plotted on normal probability paper, which immediately places T in a one-to-one correspondence with the normalized value, y, that has zero mean, 1.0 variance, and a Gaussian distribution (fig. 1). Such a transformation will yield y_0 to correspond to the initial value, T_0, and y_t to correspond to the later value, T_t.

As in a previous paper (Gringorten 1968), we assume a stochastic process to relate y_t to the earlier value, y_0, as follows:

$$y_t=\rho y_0+\sqrt{(1-\rho^2)}\eta_t \qquad (1)$$

where ρ is the correlation coefficient between y_0 and y_t separated by the time interval t (hr) and η_t is a random normal number. The process is Markov if

$$\rho=\rho_0^t \qquad (2)$$

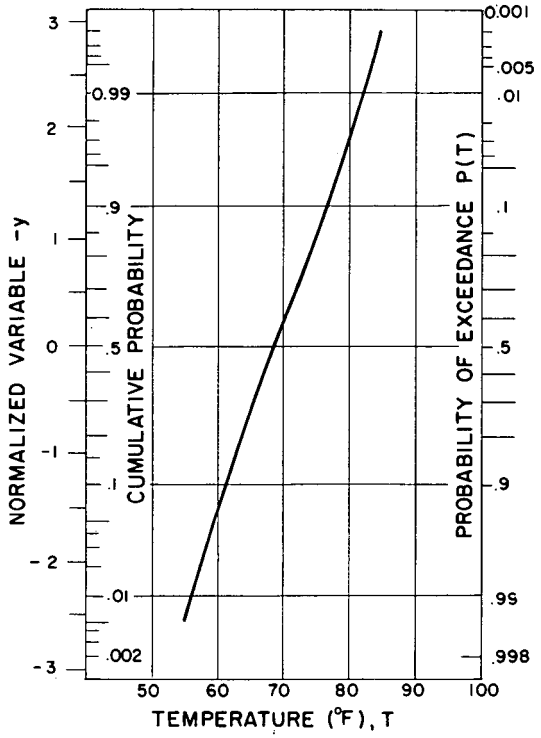

FIGURE 1.—Example of a variable, temperature (T), plotted against its cumulative frequency to give a one-to-one transformation of T into the normalized variable, y (0,1). The example is for July midnight temperature at Minneapolis, Minn.

where ρ_0 is the hour-to-hour correlation. If ρ_0 is kept constant, the process is also stationary and becomes the Markov process that is known as the "Ornstein-Uhlenbeck" process (Kendall and Buckland 1957).

It is clear from eq (1) that, for a specific value of y_0, the value of η_t will exceed a minimum, η_c, as frequently as y_t exceeds an assigned minimum, y_c. Or, where P denotes probability,

$$P(\eta_t\geq\eta_c)=P(y_t\geq y_c|y_0).$$

Hence, η_t can be denoted as $y(t|0)$, the normalized value corresponding to the conditional probability of y_t. Thus,

eq (1) can be rewritten as

$$y(t|0)=\frac{y_t-\rho y_0}{\sqrt{1-\rho^2}}. \quad (3)$$

This equation is comparable in all respects with eq (21) of the previous paper (Gringorten 1971), except that the initial condition, y_0, must be a specific value, not a category of values (i.e., $\geq y_0$). Whereas, in the previous model, the parameters λ and μ are made functions of the initial value, y_0, and persistence is given as ρ, the values y_0 and ρ enter directly into eq (3), permitting a direct and simple calculation of $y(t|0)$. Once $y(t|0)$ is obtained, a table of the normal probability integral (found in almost any text in statistics) will provide $P(y_t \geq y|y_0)$ corresponding to $y=y(t|0)$.

3. TEST AND APPLICATION

July Temperatures at Minneapolis, Minnesota

The hour-to-hour correlation had previously been set at $\rho_0=0.977$. From eq (2), the correlations, or persistence factors, between observations 3 hr apart and 15 hr apart become, respectively, $\rho=0.933$ and 0.705. From a 10-yr sample (1943–52), the upper 2-percent temperatures at midnight, 0300, and 1500 CST were reported as 81°, 78°, and 97°F, respectively. For the temperature at midnight initially equal to or higher than 81°F, the conditional frequencies of the 0300 and 1500 CST 2 percentiles (from a sample of only six cases in 10 yr) were found, by graphical interpolation, to be 0.50 and 0.42, respectively.

To use eq (3), we must represent the upper 2 percentile of the latter event by

$$y_t=2.05.$$

If the initial temperature is set successively at

$$T_0=81°,\ 82°,\ 83°,\ 84°,\ \text{and}\ 85°\text{F},$$

for which the climatic frequencies are

$$P_0=0.02,\ 0.010,\ 0.0055,\ 0.0033,\ \text{and}\ 0.0020,$$

then, from normal probability tables,

$$y_0=2.05,\ 2.33,\ 2.54,\ 2.72,\ \text{and}\ 2.88.$$

(These values are also obtainable from the left-hand scale of fig. 1.)

Placing ρ, y_0, and y_t in eq (3) gives $y(t|0)$, with which one can determine $P(T_t|T_0)$ from the normal probability tables (table 1). The temperature at the early hour of 0300 CST will equal or exceed 78°F with 35-percent probability if initially, at midnight, it is 81°F, with 63-percent probability if initially it is 82°F, and with 96-percent probability if initially it is 85°F. The afternoon temperature, 15 hr after midnight, will equal or exceed 97°F with 20-percent probability when initially it is 81°F, with 28-percent probability when initially it is 82°F, and so on. Table 1 is merely a sample of results of the application of eq (2) and (3).

TABLE 1.—*The conditional probability,* $P(T_t|T_0)$, *of the later temperature,* T_t, *given the specific initial temperature,* T_0, *at midnight for Minneapolis, Minn., in July. The later temperature is the 2 percentile for the hour* ($P=0.02 \rightarrow y_t=2.05$). *The hour-to-hour correlation is assumed to be 0.977.*

Initial midnight temperature			Later event ($P_t=0.02$)			
		Time T_t	0300 CST ≥ 78°F		1500 CST ≥97°F	
T_0	$P_0 \rightarrow y_0$		$y(t\|0)$	$\rightarrow P(T_t\|T_0)$	$y(t\|0)$	$\rightarrow P(T_t\|T_0)$
(°F)						
81	0.02	2.05	0.38	0.35	0.85	0.20
82	.010	2.33	−.34	.63	.57	.28
83	.0055	2.54	−.88	.81	.36	.36
84	.0033	2.72	−1.35	.91	.19	.42
85	.0020	2.88	−1.76	.96	.03	.49
			1943–52 samples results			
≥81				(0.50)		(0.42)

TABLE 2.—*The conditional probability,* $P(X_t|X_0)$, *of cloud cover equaled or exceeded at 2100* CST *following the specific cloud cover (nearest tenth) 12 hr earlier at 0900* CST *at Minneapolis, Minn., in January. The climatic frequencies,* P_0 *and* P_t, *are based on 1943–52 data. The hour-to-hour correlation is assumed to be 0.935.*

Initially at 0900 CST			Later cloud cover at 2100 CST (X_t) ≥1/10	≥5/10	Overcast
		P_t	0.66	0.51	0.36
		y_t	−.41	−.03	.36
(X_0)	$P_0 \rightarrow y_0$		$P(X_t\|X_0)$		
Clear sky	1.00		From 10-yr sample (0.48)	(0.19)	(0.07)
Use	0.945	−1.60	.37	.22	.12
1/10	0.89	−1.23	0.44	0.28	0.15
2/10	.78	−0.77	.53	.36	.21
3/10	.74	−.64	.56	.39	.24
4/10	.72	−.58	.57	.40	.25
5/10	.69	−.50	.58	.41	.26
6/10	.66	−.41	.60	.43	.27
7/10	.64	−.36	.61	.44	.28
8/10	.62	−.31	.62	.45	.29
9/10	.57	−.18	.64	.48	.31
Overcast	0.45		From 10-yr sample (0.77)	(0.66)	(0.53)
Use	0.225	0.76	.80	.66	.49

January Cloud Cover at Minneapolis, Minnesota

Suppose the cloud cover at 0900 CST is given to the nearest tenth (Gringorten 1971, table 6). If the hour-to-hour correlation is again assumed to be 0.935, together with the known frequencies, P_0 and P_t, then, after finding the corresponding y_0 and y_t from normal probability tables and using eq (3), the conditional probabilities of later cloud cover at 2100 CST become as shown in table 2. The probability estimate of the evening sky cover increases monotonically with the amount of the initial morning sky cover.

The only sample figures that might be usable for comparison with the estimates of table 2 are for cloud cover following an overcast or a clear sky. Since overcast is a

TABLE 3.—*The conditional probability,* $P(R_t|R_0)$, *of rainfall,* R_t, *equaled or exceeded on second day following the specific rainfall amount,* R_0, *(in.) on the first day at Boston, Mass., in January. Day-to-day correlation is assumed to be 0.21.*

			Second day rainfall (R_t)				
	First day initial rainfall		$\geq T$	≥ 0.01	≥ 0.1	≥ 0.5	≥ 1.0
		P_t	0.55	0.38	0.23	0.11	0.02
R_0	$P_0 \rightarrow y_0$	y_t	−.13	.31	.74	1.22	2.05
No rain			From 18-yr sample				
			(0.49)	(0.35)	(0.22)	(0.11)	(0.02)
Use	0.775	−0.75	.49	.32	.18	.08	.01
T	0.55	−0.13	0.54	0.37	0.19	0.10	0.02
0.01	.38	.31	.58	.40	.25	.12	.02
0.1	.23	.74	.61	.44	.27	.14	.03
0.5	.11	1.22	.66	.48	.31	.16	.03
1.0	.02	2.05	.72	.55	.37	.21	.05

category with 0.45 climatic frequency, a single value of y_0 cannot be assigned to it. But, if the value of y_0 is selected to correspond to one-half its frequency (making $y_0=0.76$ for $P_0=0.225$), the estimates of conditional probability for cloud cover $\geq 1/10$, $\geq 5/10$, and overcast are in reasonably good agreement with the sample conditional frequencies. If, for an initial category of clear sky, we make $y=-1.60$ correspond to $P=1-\frac{1}{2}(0.11)=0.945$, the results are somewhat less encouraging.

January 24-Hour Rainfall in Boston, Massachusetts

To use eq (2) and (3) on the Boston, Mass., rainfall in January (Gringorten 1971, table 11) on the conditional probability 1 day ahead, we assume $\rho_0=0.21$ and set $t=1$ (day). Then for rainfall on the first day equal to trace, 0.01 in., 0.1 in., 0.5 in., and 1.0 in., for which climatic frequencies are 0.55, 0.38, 0.23, 0.11, 0.02, respectively, the estimates of conditional probability of rain on the second day are shown in table 3. For the initial event of *no rain*, which has climatic frequency 0.45, P_0 is set at $0.55+\frac{1}{2}(0.45)=0.775$ for which $y_0=-0.75$. The resulting estimates of conditional probabilities are shown for comparison with the 18-yr sample conditional frequencies (in parentheses).

As in the previous examples, table 3 offers only a limited comparison with sampling results. It is reasonable to compare, by sample and by model, the two no-rain probabilities that are estimates. But, to satisfactorily verify the model's probability estimate of rainfall exceeding 0.01 in. or, worse, 1.0 in., following the measured amount of rainfall of the previous day, one would need a prohibitively long historical record.

4. CONCLUSIONS

Equation (3), supplemented by eq (2), is a valid and practical model for the estimate of conditional probability when the initial event can be given as a single, well-defined value. For an initial category like *overcast* or *no rain*, the model is, strictly speaking, not applicable. However, by arbitrarily assigning a probability, P_0, equal to one-half of the climatic frequency, to the initial event and inserting the corresponding normalized variable, y_0, in eq (3), the resulting estimate of conditional probability appears generally acceptable, thus obviating the need for a more elaborate procedure.

Clearly, the specific initial event, such as temperature to the nearest 1°F, cloud cover to the nearest tenth, or rainfall to the nearest 0.01 in., makes considerable difference on the conditional probability of the later threshold value. Testing this result on actual data, however, is difficult because the data sample is quickly fragmented by selecting a single initial temperature to the nearest 1°F, cloud cover to the nearest tenth, or rainfall to the nearest 0.01 in. Confidence in this model must be built upon the effectiveness of the underlying assumptions. The latter were tested in the previous work (Gringorten 1971) where larger, unfragmented samples were used for verification.

The model of eq (3) presumes a prior knowledge of persistence measured as ρ, as well as the basic climatic frequencies. To find the best values of ρ, we must use the archived data in a reversal of the problem treated in this paper and the previous paper (Gringorten 1971). But, because of the limitations of the data records, the persistence factor, ρ, should be obtained from a treatment of the data in adequately large categories, using the model of the bivariate normal distribution, as in the previous paper.

REFERENCES

Gringorten, Irving I., "Estimating Finite-Time Maxima and Minima of a Stationary Gaussian Ornstein-Uhlenbeck Process by Monte Carlo Simulation," *Journal of the American Statistical Association*, Vol. 63, No. 4, Dec. **1968**, pp. 1517–1521.

Gringorten, Irving I., "Modelling Conditional Probability," *Journal of Applied Meteorology*, Vol. 10, No. 4, Aug. **1971**, pp. 646–657.

Kendall, Maurice G., and Buckland, William R., *A Dictionary of Statistical Terms*, Oliver and Boyd, London, England, **1957**, 493 pp.

[*Received April 13, 1972; revised July 24, 1972*]